研究論文

数量化と次元縮約を伴ったファジィ c-平均法

赤木　辰伎（岡山理科大学大学院マネジメント研究科）
森　　裕一*（岡山理科大学経営学部）
黒田　正博（岡山理科大学経営学部）
飯塚　誠也（岡山大学教育推進機構）

要約:

本研究では，高次元のカテゴリカルデータを分類することを目的に，数量化と次元縮約を伴ったファジィ *c*-平均法を提案する。数量化には等質性分析を利用し，次元縮約には主成分分析の考え方を用いて，ファジィ *c*-平均法の目的関数に組み込むことで同時推定を可能とした。数値実験で，既存の拡張ファジィ *c*-平均法及びタンデムクラスタリングと性能を比較し，提案手法が次元を縮約しながら対象をよく分類することが確認できた。

キーワード:

ソフトクラスタリング，質的データ，高次元データ，同時推定，交互最小二乗法

1.　はじめに

クラスタ分析は，データの集合を部分集合（クラスタ）に分割する手法であり，基本的なデータ解析手法としてデータマイニングでも頻繁に利用されている。

このクラスタ分析には，実際の分析場面において，次の 2 つの問題が存在する。①手法の多くが数量データのみを分析の対象としていることから，カテゴリカルデータ（質的データ）の分析を前提としていない問題，②高次元データを扱う際に分類精度が落ちてしまったり，結果の解釈が難しくなったりする問題，である。これらは，マーケティング分野や心理学分野において，しばしば大きな問題となる。両分野では，アンケートをはじめとした調査において，調査項目が質的な質問から構成され，かつ，質問数（次元数）も多くなりやすいからである。したがって，本研究では高次元のカテゴリカルデータを分類することを目的とし，

* 責任著者：yuichi-mori@ous.ac.jp

問題①，②を同時に解決するクラスタリングの新しい推定方法を，クラスタリングの 1 つの手法であるファジィ *c*-平均法において提案する。

以下，2 節で，ファジィ *c*-平均法（Fuzzy *c*-means，以下，FCM）と問題①，②それぞれを個別に解決する 2 つの拡張された FCM（拡張 FCM）について述べる。3 節で，高次元のカテゴリカルデータに対して問題①，②を同時に解決する categorical reduced fuzzy *c*-means（catRFCM）を提案する。4 節で，数値実験により，提案手法の性能を確認すると同時に，実データへの適用を試みる。最後に 5 節でまとめを行う。

2.　ファジィ *c*-平均法とその拡張方法

2.1.　クラスタリングの種類とファジィクラスタリング

クラスタのへの分割方法にはいくつかの種類があり，ひとつの枠組みとして，個体がある 1 つのクラスタのみに所属するように分類するハードクラスタリングと，複数のクラスタに所属することを許して分類するソフトクラスタリング（ファジィクラスタリングともよばれる）がある。代表的なものに，前者には k-平均法（*k*-means）（MacQueen，1967），後者には FCM（Bezdek，1981；Dunn，1973）がある。両手法とも，クラスタ中心とよばれるクラスタを代表する点と帰属度（メンバシップ）とよばれるデータがクラスタへ所属するか否かを示す変数によりクラスタを分割する。k-平均法では，このメンバシップは 0 か 1 かの 2 値をとる。FCM では，0 から 1 の範囲の割合で示し，個体が属するクラスタに重複を許すように分割する。FCM は，クラスタ間に位置するような対象に対しても柔軟な解釈が可能であることから，画像解析や自然言語処理，バイオメトリクスなどの場面で広く利用されると同時に，多くの変形と応用が研究されている。

2.2.　ファジィ *c*-平均法の目的関数とアルゴリズム

p次元のユークリッド空間上の n 個の観測点を$\mathbf{X}=(\boldsymbol{x}_1,\dots,\boldsymbol{x}_i)$ で表す$(i=1,\dots,n)$。また，$n\times c$のメンバシップ行列を $\mathbf{U}=\{u_{ik}\}$，クラスタ中心行列を $\mathbf{H}=(\boldsymbol{h}_1,\dots,\boldsymbol{h}_k)$ とする$(k=1,\dots,c,\ 1\le c<n)$。

FCM は，観測点を c 個のクラスタに分類するとき，次の目的関数(1)を最小化することで推定を行う。

$$FCM(\mathbf{U},\mathbf{H}\mid\mathbf{X})=\sum_{i=1}^{n}\sum_{k=1}^{c}u_{ik}^{m}\|\boldsymbol{x}_i-\boldsymbol{h}_k\|^2 \tag{1}$$

ここで，$\|\boldsymbol{x}_i - \boldsymbol{h}_k\|$ はユークリッド距離を表すノルムであり，クラスタのまとまりを表している。m は $m > 1$ を満たすようなべき乗パラメータであり，帰属度の曖昧さ（ファジィネス）を調整するファジィ化パラメータである。$m = 1$ の場合，ハードクラスタリングである *k*-平均法と同じ目的関数となり，m の値を大きくすることでより曖昧なクラスタリングとなる。また，ファジィ分割（重なりを許した分割）のために制約

$$\sum_{k=1}^{c} u_{ik} = 1\ ,\ \ u_{ik} \geq 0 \tag{2}$$

を設ける。以下は，FCM のアルゴリズムである。

［Step1］ファジィ化パラメータm と，クラスタ数c の値を決め，制約(2)を満たしたうえでメンバシップ行列**U** を初期化する。

［Step2］(3)式によってクラスタ中心行列**H** を計算する。

$$\boldsymbol{h}_k = \frac{\sum_{i=1}^{n}(u_{ik})^m \boldsymbol{x}_i}{\sum_{i=1}^{n}(u_{ik})^m} \tag{3}$$

［Step3］(4)式によって**U** を更新する。

$$u_{ik} = \frac{\{\|\boldsymbol{x}_i - \boldsymbol{h}_c\|\}^{-\frac{1}{m-1}}}{\sum_{k=1}^{c}\{\|\boldsymbol{x}_i - \boldsymbol{h}_k\|\}^{-\frac{1}{m-1}}} \tag{4}$$

［Step4］**U** が収束していれば終了，そうでなければ，Step2 へ戻り計算を続ける。

2.3.　拡張ファジィ *c*-平均法と拡張方法

問題①，問題②ともに，それぞれ先行研究で解決が図られている。問題①のカテゴリカルデータの扱いに対しては，本多ほか（2006）がカテゴリカルデータを扱えるよう，数量化手法を取り入れた拡張 FCM（categorical fuzzy *c*-means，以下，catFCM）を提案している。これは，FCM のクラスタ中心とメンバシップの算出を繰り返すアルゴリズムにカテゴリカルデータの数量化のステップを組み込むことで，質的変数を量的変数と同じデータ空間上に布置しながら，相互の関連をよく表すクラスタ構造を導出するモデルである。具体的には，等質性分析（数量化法の 1 つで，Correspondence Analysis や数量化法 3 類と本質的に等しい解を得る）を用い，個体と質的変数のそれぞれに数量的得点を与え，カテゴリカルデータを数量データに変換しながらクラスタリングを行うものである。catFCM は，次の目的関数(5)を最小化する。

$$catFCM(\mathbf{U},\mathbf{H}\,|\,\mathbf{X})=\sum_{i=1}^{n}\sum_{k=1}^{c}u_{ik}^{m}\left(\sum_{j=1}^{p}\left\|\boldsymbol{g}_{ij}^{\top}\boldsymbol{q}_{j}-h_{kj}\right\|^{2}\right) \tag{5}$$

ここで，$\boldsymbol{g}_j$ はカテゴリカルデータXから生成された第j 変数のダミー変数行列 $\boldsymbol{G}_j=\left(\boldsymbol{g}_{1j},\boldsymbol{g}_{2j},\dots,\boldsymbol{g}_{nj}\right)^{\top}$ で $(j=1,\dots,p)$，$\boldsymbol{q}_j$ は第j 変数のカテゴリー得点であり，

$$\boldsymbol{q}_j=\left(\boldsymbol{G}_j^{\top}\left(\sum_{k=1}^{c}\mathbf{U}_k^{m}\right)\boldsymbol{G}_j\right)^{-1}\left(\sum_{k=1}^{c}h_{kj}\boldsymbol{G}_j^{\top}\mathbf{U}_k^{m}\mathbf{1}_n\right) \tag{6}$$

によって計算される。ただし，$\mathbf{U}_k$は$\mathbf{U}$のk番目の列の要素を対角にもつ対角行列$\mathbf{U}_k=diag(u_{k1},u_{k2},\dots,u_{kn})$であり，$\mathbf{1}_n$はすべての要素が 1 の$n$次元ベクトルである。(5)式は，数量化された$\boldsymbol{g}_{ij}^{\top}\boldsymbol{q}_j$を数量データ$\boldsymbol{x}_{ij}$とみなせば, (1)式の FCM の目的関数と同等と考えられる。したがって，catFCM のアルゴリズムは，通常の FCM の反復アルゴリズムに，(6)式のカテゴリー得点を求める数量化ステップを追加することで成立する。

問題②の高次元データの扱いに対しては，西田（2010）が主成分分析による次元縮約を取り入れた拡張 FCM（reduced fuzzy *c*-means，以下，RFCM）を提案している。これは，クラスタリングそのものとクラスタリングによって得られたクラスタ中心座標の次元縮約を同時に推定するモデルである。RFCM は，次の目的関数(7)の最小化を行う。なお，推定すべきパラメータは$\mathbf{U}$，$\mathbf{F}$，$\mathbf{A}$と 3 つ存在するため，一度に推定することは難しい。そこで，最小二乗基準の意味でそれぞれの制約を満たしながら，繰り返し推定値を求めるデータ変換（同時推定）を行う交互最小二乗法（森ほか，2017 などを参照）によって 3 つのパラメータを推定する。

$$RFCM(\mathbf{U},\mathbf{F},\mathbf{A}\,|\,\mathbf{X})=\sum_{i=1}^{n}\sum_{k=1}^{c}u_{ik}^{m}\|\boldsymbol{x}_i-\mathbf{A}\boldsymbol{f}_k\|^2 \tag{7}$$

ここで，$\mathbf{A}$ は$p\times r$の主成分負荷行列であり，$\boldsymbol{f}_k$はクラスタ中心における$c\times r$の主成分得点行列$\mathbf{F}$のk 番目の行である$(1\le r<p)$。$\mathbf{A}$ には，制約

$$\mathbf{A}^{\top}\mathbf{A}=\mathbf{I}_M \tag{8}$$

があり，

$$\mathbf{A}=\mathbf{K}\mathbf{L}^{\top} \tag{9}$$

で更新される。ただし，$\mathbf{K}$，$\mathbf{L}$ は，それぞれ $\mathbf{X}^{\top}\mathbf{U}\mathbf{F}$ の特異値分解

$$\mathbf{X}^{\top}\mathbf{U}\mathbf{F}=\mathbf{K}\boldsymbol{\Delta}\mathbf{L}^{\top}$$

の左特異ベクトル，右特異ベクトルである。(7)式を見ると，catFCM 同様，$\mathbf{A}\boldsymbol{f}_k$をクラスタ中心$\boldsymbol{h}_k$とみなせば，(1)式の FCM の目的関数と同等である。したがって，RFCM のアルゴ

リズムも通常の FCM の反復アルゴリズムに，(7)式の$\mathbf{A}\boldsymbol{f}_k$を求める次元縮約のステップを追加することで成立する。

ハードクラスタリングにおいても，問題①，②それぞれに対応した手法が提案されている。カテゴリカルデータに対しては，Van Buuren & Heiser（1989）の GROUPALS，高次元データに対しては，De Soete & Carroll（1994）の Reduced k-means clustering などがある。

以上のように，カテゴリカルデータの問題①と高次元データの問題②については，各問題に個別に対応したクラスタリングが研究されている。FCM においても上記のとおり，catFCM や RFCM などの拡張 FCM が既に提案されている。しかしながら，高次元のカテゴリカルデータを分類することに対しては，既存の手法では，カテゴリカルデータの処理と高次元の処理を別々に行わなければならない。これでは，データに内在している情報の損失が懸念されるため，この 2 つの処理（問題①，②の解決）を同時に行うことが望ましい。すなわち，数量化と次元縮約の両方を取り入れた FCM がほとんど研究されていないことから，拡張 FCM である catFCM と RFCM の拡張方法の両方を取り入れて，数量化と次元縮約を同時に行う FCM を新たに提案する。

3. 数量化と次元縮約を伴ったファジィ *c*-平均法

3.1. 数量化と次元縮約の同時推定

カテゴリカル変数からなるデータ$\mathbf{X}$が与えられたとする。数量化と次元縮約の両方を取り入れた FCM の目的関数を考えるにあたり，FCM に，catFCM，RFCM の目的関数を組み合わせる。具体的には，FCM の目的関数である(1)式をベースとして，まず(5)式を利用し，(1)式の$\boldsymbol{x}_i$を数量化された$\boldsymbol{g}_{ij}^{\top}\boldsymbol{q}_j$に置き換える。次に，(7)式を利用して，(1)式の$\boldsymbol{h}_k$を主成分座標上の$\mathbf{A}\boldsymbol{f}_k$に置き換える。このようにして，提案手法である catRFCM の目的関数を以下のように定める。

$$catRFCM(\mathbf{U},\mathbf{H},\mathbf{A} \mid \mathbf{X}) = \sum_{i=1}^{n}\sum_{k=1}^{c} u_{ik}^{m}\left\|\boldsymbol{g}_{ij}^{\top}\boldsymbol{q}_j - \mathbf{A}\boldsymbol{f}_k\right\|^2 \tag{10}$$

catRFCM は，この(10)式を最小化する。つまり，提案手法は，数量化された個体と主成分座標上のクラスタ中心との距離の二乗和が最小となるようにクラスタリングを行っているといえる。

3.2. catRFCM のアルゴリズム

提案手法 catRFCM のアルゴリズムを以下に示す。

［Step1］ファジィ化パラメータm，クラスタ数c，主成分数rの値を決め，制約(2)，(8)を満たしたうえで，メンバシップ行列$\mathbf{U}$，クラスタ中心行列$\mathbf{H}$，主成分負荷行列$\mathbf{A}$を初期化する。

［Step2］(6)式によって第j変数のカテゴリー得点$\boldsymbol{q}_j$を求め，以下，$\boldsymbol{g}_{ij}^{\top}\boldsymbol{q}_j$を$\boldsymbol{x}_{ij}$とおく。

［Step3］(3)式によって$\mathbf{H}$を計算し，(11)式を用いて主成分得点行列$\mathbf{F}$を更新する。

$$\mathbf{F} = \mathbf{HA} \tag{11}$$

［Step4］(9)式によって$\mathbf{A}$を更新する。

［Step5］(4)式によって$\mathbf{U}$を更新する。

［Step6］すべてのパラメータ（$\mathbf{U}$，$\mathbf{H}$，$\mathbf{A}$）が収束していれば終了，そうでなければ，Step2へ戻り計算を続ける。

4. 数値実験

2 種類の数値実験を行う。1 つは，人工データを用意し，指標を用いて性能評価を行うもので，もう 1 つは，実データへの適用である。両実験では，扱う手法に 2 つのタンデムクラスタリング（2 種類以上の分析手法を別々に行う，二段階手続きのクラスタリング手法），tandem1 と tandem2 を追加する。tandem1 では最初にデータの数量化を行い，次に FCM を用いてクラスタリングを行う。tandem2 では最初にデータの数量化を行い，次に RFCM を用いてクラスタリングを行う。したがって，数値実験では，通常の FCM，拡張した catFCM，RFCM，タンデムクラスタリングの tandem1，tandem2 の 5 つの手法と提案手法の catRFCM で比較を行う。

4.1. 人工データによる数値実験

人工データは，R パッケージ `MixSim` を使用し，指定したクラスタ数からなる量的データを生成した後，それらの量的変数をすべて 5 段階からなる質的変数に変換して，全変数がカテゴリカルである人工データを作成した。変換にあたっては，選択肢 1～5 の個体数がほぼ同じになるようにした。このとき，`MixSim` で生成された各個体は，所属しているクラスタ，つまり正解のクラスタ番号が明らかになっている。人工データは，$n = 100$，$p = 12$，$c = 4$として 100 セット生成し，クラスタリングにおいては，ファジィ化パラメータm

は$m = 1.5$とし，主成分数rは$r = 4$（累積寄与率は 0.67~0.76 の間）としてシミュレーションを行う。

性能評価は, 3 つの指標, Fuzziness Performance Index（FPI）, Normalized Classification Entropy（NCE），Misclassification Probability（MCP）を用いる。FPI は，

$$\mathrm{FPI} = \frac{1 - (c \times \mathrm{PC} - 1)}{(c - 1)}$$

で求める（Roubens，1982）。PC は，Partition Coefficient（Bezdek，1974）のことで，

$$\mathrm{PC} = \frac{1}{n}\sum_{i=1}^{n}\sum_{k=1}^{c} u_{ki}^2$$

である。NCE は，

$$\mathrm{NCE} = \frac{\mathrm{PE}}{\log c}$$

で求める（Roubens，1982）。PE は，Partition Entropy（Bezdek，1974）のことで，

$$\mathrm{PE} = -\frac{1}{n}\sum_{i=1}^{n}\sum_{k=1}^{c} u_{ki} \log u_{ki}$$

である。FPI と NCE は，クラスタの分離状態，つまりクラスタが互いにどの程度分離されているかを示す指標であり，これらの値が小さいほど，クラスタが明確に分離されていることを表す。MCP は，

$$\mathrm{MCP} = \frac{\mathrm{MC}}{n}$$

で求める。MC は，推定によって不正解のクラスタに所属された個体の数である。MCP は，生成時と推定時の各個体のクラスタ番号を比較し，推定によって誤ったクラスタに分類された確率を示す指標である。そのため，値が小さいほど，クラスタリングの精度がよいことを表す。

表 4.1 は，100 回のシミュレーション結果から，各手法の FPI，NCE，MCP の平均値をまとめたものである。また，図 4.1 は，100 回のシミュレーションによる FPI，NCE，MCP を手法ごとに箱ひげ図にしたものである。まず，FPI と NCE の値に注目すると，表 4.1 より，提案手法 catRFCM の値は，すべての手法の中で最も小さく，図 4.1 の箱ひげ図からも，提案手法は，他の 5 つの手法と比較して，小さな値が多く布置していることがわかる。また，提案手法の次には catFCM の値が小さく，カテゴリカルデータには数量化を行うことが重要

表 4.1　各手法における推定結果の比較（100 回のシミュレーション結果より平均値を算出）

手　法	FPI	NCE	MCP
catRFCM（提案手法）	0.44	0.48	4.5%
FCM	0.53	0.56	3.9%
catFCM	0.46	0.49	6.2%
RFCM	0.56	0.59	3.6%
tandem1	0.51	0.53	10.6%
tandem2	0.54	0.56	11.3%

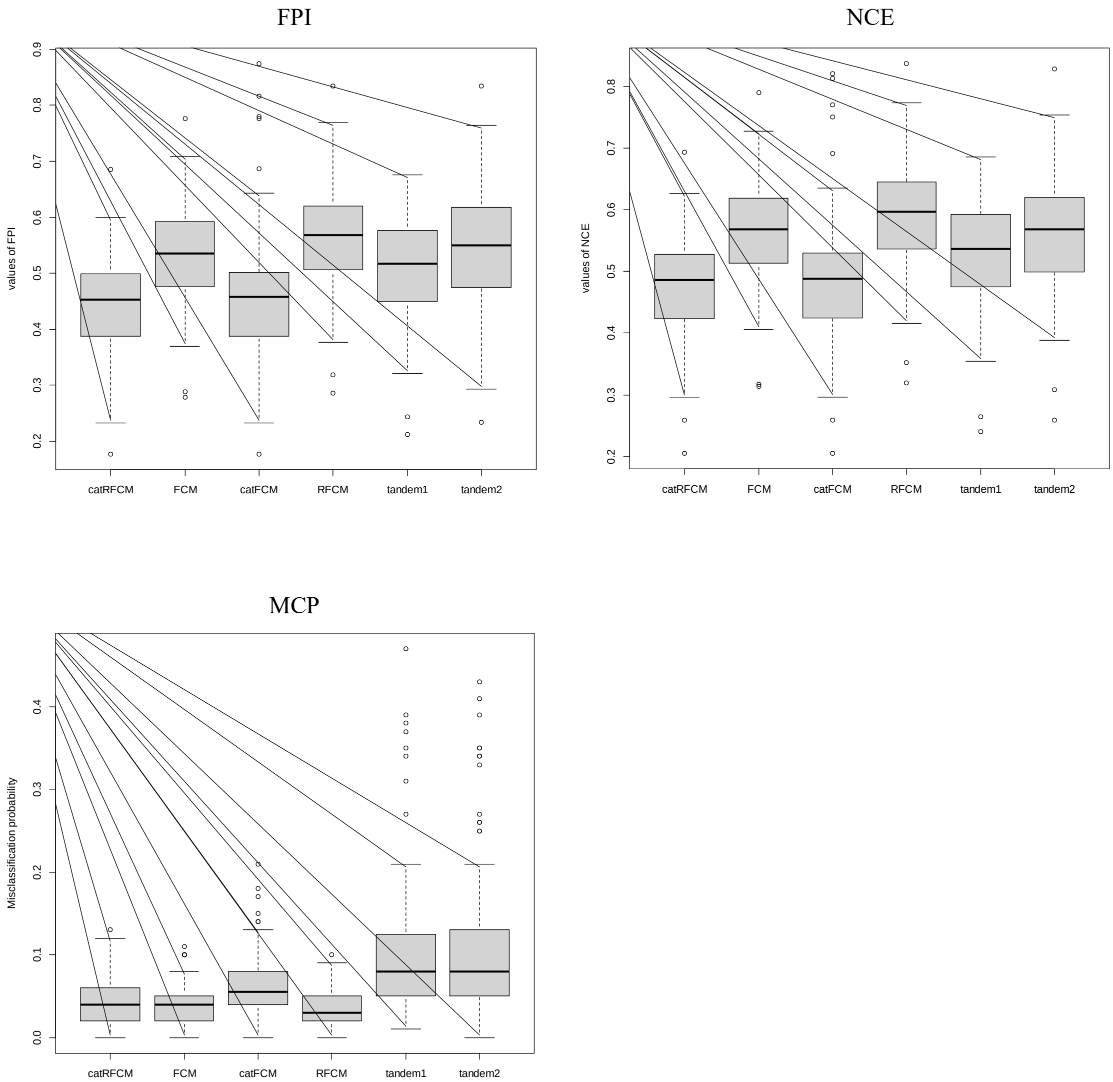

図 4.1　手法ごとの 100 回のシミュレーションによる FPI，NCE，MCP の箱ひげ図（各図の左から，catRFCM，FCM，catFCM，RFCM，tandem1，tandem2）

であることがうかがえる。次に，表 4.1 の MCP を見ると，4 つの手法（catRFCM，FCM，catFCM，RFCM）は，その値がすべて 3%から 6%の間に収まっている。このことから，これら 4 つの手法は，生成時の各個体の所属クラスタをおおむね正しく再現できているといえる。一方，tandem1，tandem2 の MCP は，どちらも 10%を超えており，図 4.1 の MCP の箱ひげ図からも，タンデムクラスタリングの分類精度は高くないことがわかる。タンデムクラスタリングは，数量化や次元縮約をクラスタリングとは別に事前に推定するため，FCM での推定前にクラスタリングに必要な情報が落ち，分類精度が悪くなったと考えられる。

以上の結果から，通常の FCM と拡張 FCM（catRFCM，catFCM，RFCM）では，FPI と NCE の観点から提案手法が最も明確にクラスタを分離するといえる。誤分類率の観点からも総合的に判断して，提案手法の性能は十分なものであるといえる。

4.2. 実データへの適用

使用するデータは，ファッションに関するアンケート調査（Ohyabu et al., 2019）である。ここで，マーケティング，特にファッションの分野では，消費者にいくつかのタイプがいることが知られている。例えば，ブランドとショップスタッフの価値の関係において，以下のようなタイプがいる。

- 好きなブランドがあり，ショップスタッフの価値を重視している人
- 少し好きなブランドがあり，ショップスタッフの価値はあまり重視していない人
- ブランド，ショップスタッフのどちらにもあまり価値を見出さない人

このアンケート調査は，ファッションが好きな人やファッションにこだわりがある人，洋服を購入する際，特定のショップ（例えば，Beams などのセレクトショップ）で購入することが多い人を対象に，「関与」，「顧客エンゲージメント」，「価値」について，ファッション・ブランド・ショップスタッフの 3 つの観点から質問し，回答を求めたものである。調査データは，調査対象者が 824 名，質問項目が 85 項目である。回答選択肢は，質問により多少変化するものの，1.とてもあてはまる，2.少しあてはまる，3.どちらでもない，4.あまりあてはまらない，5.まったくあてはまらない，のように，選択肢 1 が肯定的で，5 に近づくにつれて否定的な回答になるように設定されている。このデータから，「価値」に関する 34 個の質問項目に対して，無作為に抽出した 40 名をデータとして用いる。質問項目は，1～11 がファッションの観点からの質問，12～23 がブランドの観点からの質問，23～34 がショップスタッフの観点からの質問である。したがって，実験で使用するデータは，$n = 40$，$p = 34$で，

推定においては, FPI と NCE の最小値が 3 となったことからクラスタ数を 3 $(=c)$ とした。また，R の prcomp 関数を用いて，元のカテゴリカルデータに主成分分析を適用し，累積寄与率が 0.6 を超える主成分数$r=2$を採用した。ファジィ化パラメータmは$m=2$である。

図 4.2 の左上は，提案手法の推定結果から，各個体の主成分得点とクラスタ中心をプロットした散布図である。図 4.2 から，提案手法のクラスタリング結果は，クラスタ 1（c1），クラスタ 2（c2），クラスタ 3（c3）をそれぞれ中心として，各個体をよく分類しているのが見て取れる。また，図 4.2 の残りの散布図は，FCM，catFCM，RFCM，tandem1，tandem2 の 5 つの手法の推定結果の散布図であり，これらと比較すると，提案手法のクラスタリング結果が最も明確にクラスタを分離させており，個体とクラスタの構造や関係が他の手法よりよく見て取れることがわかる。

また，図 4.3 は，提案手法によって分けられた 3 つのクラスタ（c1，c2，c3）において，選択肢ごとの度数を質問ごとに積み上げ棒グラフに表したものである。図 4.3 から，c1 は，ブランドの価値に関する質問（12〜23）において肯定（青色）からやや肯定（水色）の回答が半数で，ショップスタッフの価値に関する質問（23〜34）においては，他クラスタと比較して否定（赤色）ややや否定（橙色）の回答が目立つ。このことから，c1 は，少し好きなブランドがあり，ショップスタッフの価値はあまり重視していない消費者たちからなるクラスタと解釈できる。一方，c2 は，すべての観点で肯定的な回答が目立つ。よって，c2 は，好きなブランドがあり，ショップスタッフの価値を重視している消費者からなるクラスタと解釈できる。最後に，c3 は，ブランドとショップスタッフ両方の質問において肯定が少なく，否定的な意見がいずれの質問でも確認できることから，ブランド，ショップスタッフのどちらにもあまり価値を見出さない消費者たちからなるクラスタと解釈できる。

以上のように，人工データと実データへの適用により，本手法は，実際の高次元カテゴリカルデータに対して，他手法と比較して明確にクラスタの構造を表したクラスタリングを行い，そのクラスタは，実際の場面をよく表現しているものであったといえる。特に，次元縮約により，図 4.2 のような散布図による可視化が容易となることは，マーケティング分野などの実際の分析場面で大きく役立つといえる。例えば，今回のアンケート調査においては，提案手法を用いた可視化により各クラスタの特徴を把握することで，クラスタごと，消費者に適切なファッション広告・販促施策を投じるといったことが可能になる。他にも，クラスタ間にある個体を視覚的に把握することで，その個体のメンバシップ値に応じて，2 つのクラスタの特徴をもつ消費者には複数の広告をうつなど，柔軟な解釈と応用が可能である。

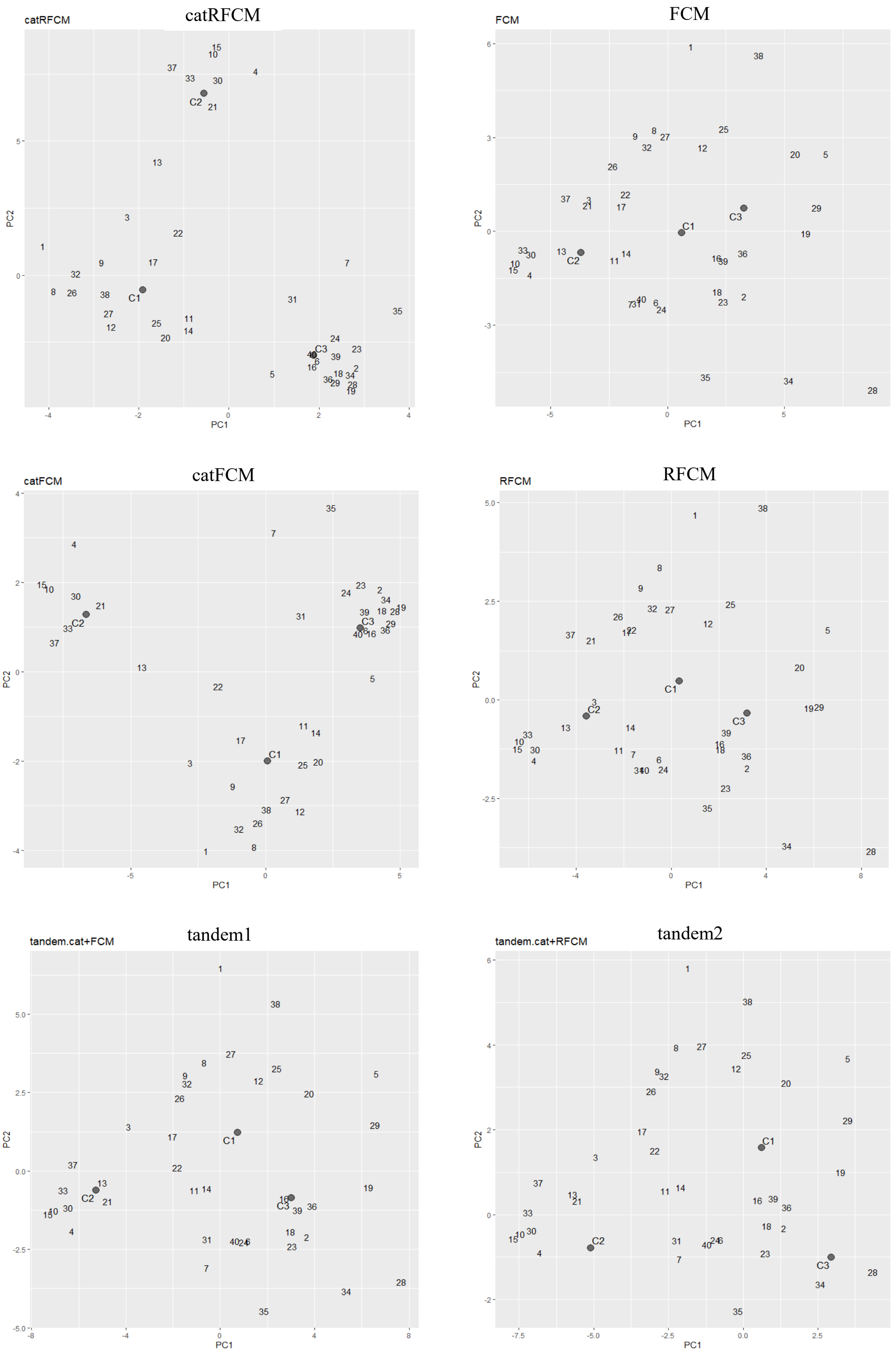

図 4.2　アンケート調査のクラスタリング結果の比較

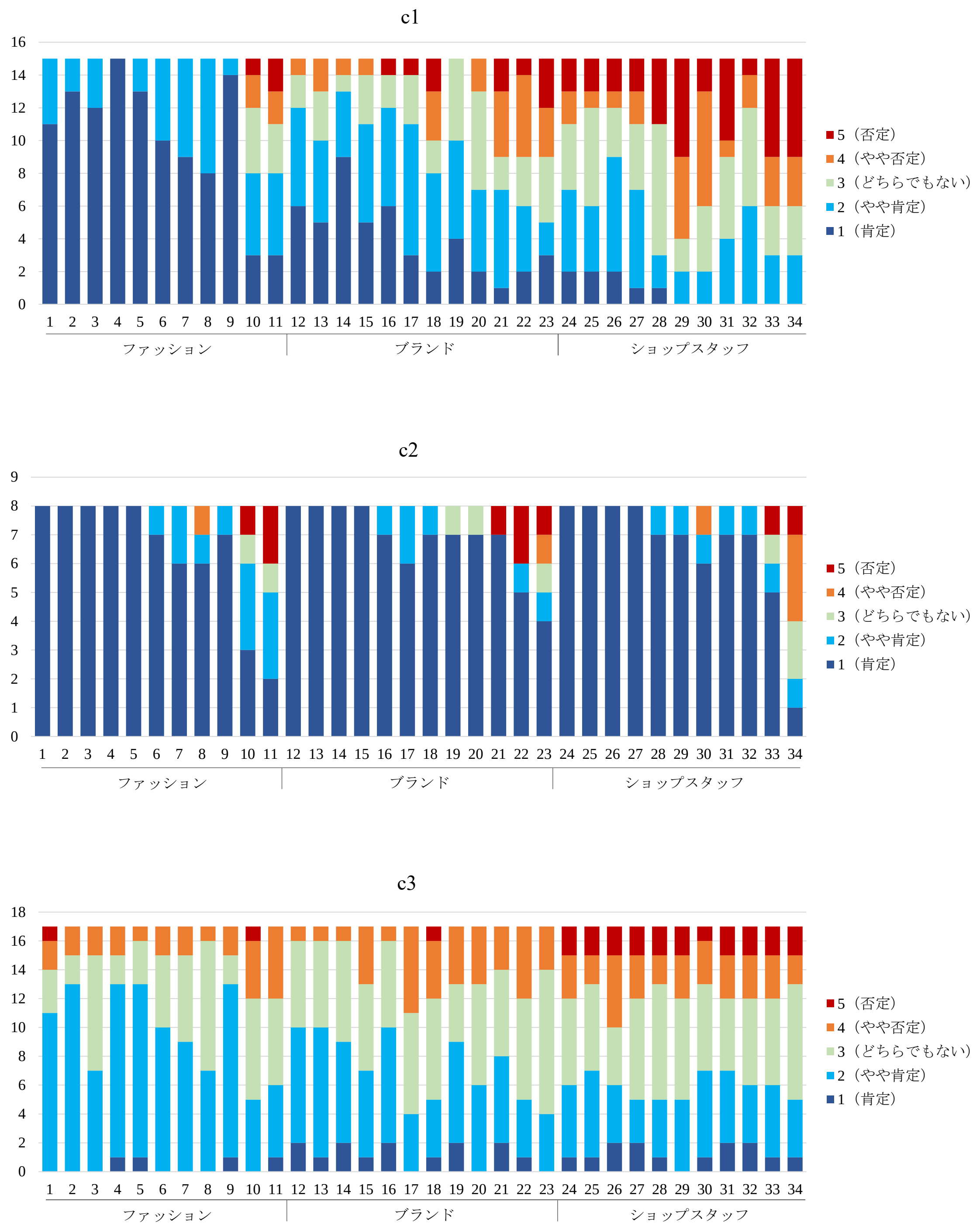

図 4.3　catRFCM により分けられたクラスタ c1，c2，c3 における各質問に対する選択肢の選択数（縦軸：回答数（=個体数），横軸：質問項目，棒内の色：選択肢の種類（紺：1，青：2，薄緑：3，橙：4，赤：5））

5. 結　論

本研究では，クラスタリングで遭遇する 2 つの問題（問題①カテゴリカルデータの処理，問題②高次元データの処理）を同時に解決するため，数量化と次元縮約の両方を取り入れたファジィ *c*-平均法（categorical reduced fuzzy *c*-means，catRFCM）を提案し，高次元のカテゴリカルデータを分類することを試みた。数量化には等質性分析を利用し，次元縮約には主成分分析の考え方を用いて，それらをファジィ *c*-平均法の目的関数とアルゴリズムに組み込むことで，数量化，次元縮約，ファジィクラスタリングの同時推定を可能とした。数値実験により，提案手法を通常のファジィ *c*-平均法，既存の拡張ファジィ *c*-平均法及びタンデムクラスタリングと比較し，高次元のカテゴリカルデータに対する性能を確認した。その結果，高次元のカテゴリカルデータに対して，提案手法において以下の利点が明らかになった。

- 提案手法は，通常のファジィ *c*-平均法，既存の拡張ファジィ *c*-平均法と比較して，クラスタをより明確に分離する。
- 提案手法を含む同時推定法は，タンデムクラスタリングと比較して分類精度が高い。
- 提案手法は，実際の場面に解釈可能な分析結果を導出する。

これにより，これまで 2 つの問題を個別にしか解決できなかったことが 1 つの目的関数の下でパラメータが推定できるようになったこと，個別の推定に比べ，推定結果が元のデータのもつ構造をよく反映できていることから，実用に向けた手法の基礎が提案できたと考える。

今後の課題として，他の指標に基づいて，引き続き提案手法の性能評価をすることがあげられる。特に，ソフトクラスタリングの観点から，クラスタの分離について，メンバシップ値を考慮した指標による性能評価が課題となる。また，他の実データに適用し，さまざまな場面での利用価値を検討することが必要と考える。

謝　辞

本研究は JSPS 科研費 JP21K11799，JP21K11800，JP22K02714 の助成を受けたものです。

参考文献

Bezdek, J. C. (1974). Cluster validity with fuzzy set. *Journal of Cybernetics*, **3**, 58-72.

Bezdek, J. C. (1981). *Pattern Recognition with Fuzzy Objective Function Algorithms*, Springer.

De Soete, G., & Carroll, J. D. (1994). K-means clustering in a low-dimensional Euclidean space. In

Diday, E., Lechevallier, Y., Schader, M., Bertrand, P., and Burtschy, B (Eds.). *New Approaches in Classification and Data Analysis*, Springer, Heidelberg, 212-219.

Dunn, J. C. (1973). A Fuzzy Relative of the ISODATA Process and Its Use in Detecting Compact Well-Separated Clusters. *Journal of Cybernetics*, **3**(3), 32-57.

MacQueen, J. B. (1967). Some Methods for classification and Analysis of Multivariate Observations. *Proceedings of 5th Berkeley Symposium on Mathematical Statistics and Probability*, **1**, 281-297.

Ohyabu, A., Kuroda, M., Seino, S., & Zhang, J. (2019). Exploring interplay among consumer engagements with multiple objects. The 10th Years Naples Forum on Service-Service Dominant Logic, http://www.naplesforumonservice.it/uploads/files/2018/BookOfAbstract/BoA_Ohyabu-Kuroda-Seino-Zhang.pdf

Roubens, M. (1982). Fuzzy clustering algorithms and their cluster validity. *European Journal of Operational Research*, **10**, 294-301.

van Buuren, S., & Heiser, W. J. (1989). Clustering *N* objects into *K* groups under optimal scaling of variables. *Psychometrika*, **54**, 699-706.

西田豊 (2010). 曖昧さを表現する概念形成モデル. 知能と情報（日本知能情報ファジィ学会誌）, **22**(4), 434-442.

本多克宏・上杉亮・市橋秀友 (2006). 混合データベースの FCM クラスタリング. 知能と情報（日本知能情報ファジィ学会誌）, **18**(4), 598-608.

森裕一・黒田正博・足立浩平 (2017).『最小二乗法・交互最小二乗法』共立出版.